DOI：10.13998/j.cnki.issn1002-1248.23-0200

畜牧兽医基因组学领域技术空白中外对比研究

吴　蕾[1,2]，李小杰[3]，丁　倩[1,2]，孙　巍[1,2]，周正奎[4]
（1. 中国农业科学院农业信息研究所，北京　100081；2. 农业农村部 农业大数据重点实验室，北京　100081；
3. 中国农业科学院深圳农业基因组研究所，深圳　518120；4. 中国农业科学院北京畜牧兽医研究所，北京　100193）

摘　要：[目的 / 意义]为了挖掘中国在农业重点领域的技术空白，并预测空白点的未来发展趋势，为科技管理决策者提供有效的科技发展技术机会咨询建议。[方法 / 过程]首先，使用关键句嵌入方法和句向量聚类方法，对论文和专利的摘要信息进行挖掘；然后进行主题聚类对比分析，发现技术空白；其次，构建语义相似性网络和分类相似性网络，发现容易与空白点形成交叉融合的主题方向。[结果 / 结论]在畜牧兽医领域对基因组学技术进行了实证分析。结果表明，该方法能够发现技术空白，并结合专家分析，可以对畜牧兽医领域基因组学技术进行发展现状解读和未来趋势预测，并为中国畜牧兽医领域基因组学技术智库咨询提供方法和数据支撑。

关键词：技术空白发现；关键句抽取；句向量聚类；基因组学；知识产权
中图分类号：G255.51　　**文献标识码：**A　　**文章编号：**1002-1248（2023）08-0088-10

引用本文：吴蕾，李小杰，丁倩，等. 畜牧兽医基因组学领域技术空白中外对比研究[J]. 农业图书情报学报，2023，35（8）：89-97.

1 引　言

科技创新是指科技在发展进步过程中发生的变化。科技管理决策者需要了解科技创新发展的进程和动态，掌握甚至预测这种科技创新契机的存在，从而利用有限的资源支持和促进科技进步。因此，深入了解科技发展趋势，通过有效方式识别领域和行业潜在的技术空白至关重要。

技术空白指一个技术领域中尚未开发，但具有很强技术创新潜力的部分。体现为在现有专利中还未有技术布局的概念或某些概念的组合[1]。目前，针对技术空白的相关研究主要集中在技术空白识别[1,2]。这些研究没有考察技术空白未来的发展趋势，因此无法为科技管理决策者提供进一步的咨询建议。本文在技术空白发现基础上，进一步在国内外众多成熟的研究和技术方向中锁定与空白点有相似基础研究分类的成熟方向。成熟方向指在论文和专利中都已形成独立主题，

收稿日期：2023-03-30
基金项目：国家社科基金青年项目“基于图模型的农业领域多源知识迁移研究”（18CTQ028）；国家重点研发计划项目“智能化情报分析软件工具研发”子课题“领域演化分析工具研发”（2022YFF0711904）
作者简介：吴蕾（1985-），女，助理研究员，研究方向为文本挖掘与情报分析。共同第一作者李小杰（1982-），男，副研究员，研究方向为农业生物组学发展战略。丁倩（1987-），女，助理研究员，研究方向为农业信息管理。孙巍（1978-），女，研究员，研究方向为技术预测预见。周正奎（1979-），男，研究员，研究方向为动物遗传育种与繁殖

即成熟方向在基础研究和技术应用领域已经形成完整链条。由于与技术空白在基础研究领域的相似性，因此成熟方向可能在实验条件、实验方法等领域提供参考价值，进而与技术空白点在基础研究领域形成交叉融合，并促进空白点的研究基础进一步发展，有利于技术空白的发展和填补。本文更进一步提取了与成熟方向拥有相似基础研究的其他主题方向，这些方向也可能为技术空白的发展提供些许参考。另外，成熟技术在应用场景、技术功效等方向也可能为技术空白指引方向。当某项技术空白只存在于中国时，通过挖掘与成熟技术应用领域相似的国外技术，可以为中国技术空白的发展和填补提供参考。当某项技术空白限制全球技术发展时，在基础研究领域寻求支撑可能是一条出路。同样的，本文在挖掘成熟方向技术应用领域的同时，也提取了与成熟方向拥有相似应用领域的其他技术方向，相同技术部类下相似技术方向之间形成的相似性技术融合也可能为技术空白的发展提供些许参考[3]。

通过对比中外畜牧兽医领域基因组学技术分布，挖掘中国技术空白点，并预测空白点在基础研究领域和技术应用领域的未来走向，能帮助科技管理决策者把握科学技术发展全流程和新动向，对科学研究和技术发展具有指导意义。

2　相关工作

随着科学与技术的快速发展，技术改进越来越显现出递归性特征。即某种新技术不能凭空而生，而是有迹可循。目前技术空白发现方法主要包括两种：第一种是基于专利地图的方法。这类方法首先需要构建关键词关联矩阵或者关键词向量，然后通过主成分分析[4,5]、自组织映射[6]、生成式拓扑映射[1]等方法进行映射，最终绘制成专利地图，并从中发现技术空白点。第二种是利用技术功效矩阵的方法。这类方法从技术和功效两个维度来分析当前某个领域的专利，以此来寻找未被研发出来的新领域[7]。这类方法需要对专利数据逐篇标引技术类型和功效，以技术类型为纵轴、功效为横轴绘制表格。这类方法的缺点是功效的定义有时很难区隔，同时提升质量和降低成本已能够涵盖大部分功效。本文使用关键句而不是关键词来表征论文和专利，能够提供更多语义信息。另外，通过聚类方法可以避免逐篇标引带来的工作量和主观误差。通过对比中外论文和专利的主题聚类分布，能够发现中国技术空白点，并为后续发展预测提供支撑。

为了衡量技术之间的关系，并借此寻找可能促进技术空白点发展的主题方向，研究者主要采用共类分析识别方法，使用的分类属性包括：标准产业分类代码（SIC）[8]、专利分类代码[9]、关键词等[10]。文章[11]利用叙词表对专利 IPC 号进行技术领域归类，然后构建领域共现网络，并利用余弦相似度计算 IPC 分类颗粒度的技术领域融合度。但是由于这类技术分类颗粒度较粗，因此揭示的技术领域融合方向也较为宽泛。本文利用论文的 WOS 分类和专利的 IPC 分类获取各聚类中论文和专利的基础研究分类分布和技术应用分类分布。然后通过计算分类分布向量的相似度，衡量论文或专利中各聚类主题的分类相似性。从而避免了分类层级颗粒度较粗的问题。另外，本文同时使用各聚类的嵌入向量相似度衡量各聚类主题的语义相似性。

3　研究方法

本文使用关键句嵌入方法挖掘关键句，并转化成句向量进行聚类。通过对聚类结果进行对比分析及解读，挖掘技术空白。通过相似性分析，发现容易与技术空白形成交叉融合的主题方向，最终完成技术空白发现与预测，达到咨询建议的目的。

3.1　关键句嵌入及聚类

首先，采用 TextRank 算法分别从论文和专利的摘要信息中抽取关键句。关键句是对文本集合的抽取或凝练。借助文本的语义关键信息，可以减少领域专家对聚类结果标注的工作量和时间[12]。TextRank 算法可用于进行无监督关键句提取，其将摘要中的句子作为网络中的节点，将句子与句子之间的共现关系表示成

DOI：10.13998/j.cnki.issn1002-1248.23-0200

网络中节点之间的边，将句子对之间的共现相似度作为边的权重。TextRank 算法常被用来从给定的文本中抽取关键词、关键词组和关键句[13]。

然后，将 TextRank 算法提取的关键句作为 Sentence-BERT 算法的输入，得到关键句嵌入。与 word2vec[14]、FastText[15]等词嵌入算法类似，Sentence-BERT 算法[16]可以将文本表示成数字向量，为后续的学习任务提供便利。但是由于关键句比关键词提供了更多的语义信息，因此可以更加便于领域专家解读每个聚类的主题含义。与 SkipThought[17]、Quick-Thoughts Vectors[18]等其他句子嵌入算法相比，Sentence-BERT 算法不但能够通过引入从大型数据集中预训练得到的通用句子来提升算法效率和泛化性能。同时，还能够更加快速地计算句子相似度，并生成关键句向量。另外，Sentence-BERT 算法在 BERT 算法的基础上使用了孪生网络和三元组网络，生成具有语义的句向量，并使用池化层固定句向量的长度，而不是使用每个 token 的上下文表示。与 BERT 算法相比，Sentence-BERT 算法拥有更高的运算效率。

其次，本文使用 K-means 聚类算法[19]和基于密度的噪声应用空间聚类算法（Density-Based Spatial Clustering of Applications with Noise，DBSCAN）[20]进行了对比实验，构建了基于关键句嵌入的中外论文和专利聚类。K-means 算法是一种常用的主题聚类算法，其保证每个聚类内部的句向量间距离尽可能小，同时保证聚类间的句向量距离尽可能大。文章[21]利用 K-means 算法对 LDA 和加权 Word2Vec 词向量的输出结果进行主题聚类。文章[22]采用基于 doc2vec 的 K-means 聚类分析。尽管 K-means 聚类方法简单、有效，但是必须预先设定聚类数量。对比发现，DBSCAN 算法的优点就是不需要指定集群的数量，但是其结果准确性更加依赖数据结构特点[23]。实验结果表明，K-means 方法可以通过人工调整聚类参数挖掘出更多有意义的主题类别。由于专利技术领域的数据量较少，因此 DBSCAN 方法仅能够完成部分数据聚类，但是仍存在大量数据没有被划分到任何聚类中。另外，K-means 方法还有一个优点，就是计算成本低，其时间复杂度为 $O(n)$（其中 n 是数据个数），且易于扩展。通过对比，DBSCAN 方法虽然无需人工参与调参，但是其聚类过程中需要执行多次距离计算，从而导致效率低下，时间复杂度为 $O(n^2)$。因此将 DBSCAN 方法应用到本文研究中是受限制的，而 K-means 方法更适合本研究。

最后，领域专家借助聚类的关键句及关键词 /IPC 分类信息对主题进行语义标注，并通过对比分析，探测技术空白。

3.2 技术空白发现及预测

考虑到科技管理决策者需要站在国家的角度把握科学技术机会的投入。因此本文从国家维度进行中外科技主题分布对比研究，并分析中国的技术空白点，有助于辅助科技管理决策者了解科技发展态势，从而进行科技政策制定。

首先，本文对中外论文和专利的聚类分布进行对比分析。通过判断在专利中是否具备技术应用，确定该聚类主题方向是否属于技术空白。然后，使用中外论文和专利摘要关键句嵌入向量聚类均值表征各个聚类的技术语义分布；使用各个聚类中论文或专利所属学科或 IPC 分类向量表示技术的基础研究分布或技术应用分布。其中分类向量的长度由所有聚类中可能出现的学科或 IPC 总数决定，向量中某个维度的数值表示该聚类里属于当前维度对应分类的论文或专利数量。

具体来看，为了探究论文和专利中主题的语义相似性，并进一步锁定成熟方向，本文首先采用 Spearman 相关系数衡量中、外论文和专利摘要聚类的语义相似性，并构建了中、外论文和专利的语义相似性网络。该网络主要描述了多源数据之间的主题相似性关系。文章[24]的研究表明针对词 / 句向量，相似度和相关性测量结果在统计意义上是相通的。另外，对于词 / 句向量，基于排序的相关性计算方法要比其他基于数值的相似度和相关性算法更加鲁棒。因此，本文使用 Spearman 相关系数衡量两个聚类中心向量之间的相似性（公式 1）。

$$\rho_X = \frac{\sum\left(R_X - \overline{R_X}\right)\left(R_{X'} - \overline{R_{X'}}\right)}{\sqrt{\sum\left(R_X - \overline{R_X}\right)^2 \sum\left(R_{X'} - \overline{R_{X'}}\right)^2}} \qquad (1)$$

X 和 X' 分别表示两个聚类的中心向量，即该聚类中所有论文或专利的摘要句向量的平均值。然后，将 X 和 X'分别从小到大排序编秩。R_X 和 $R_{X'}$ 分别表示两个秩次。ρ_X 表示两个聚类的语义相似性。

然后，使用 Spearman 相关系数衡量中外论文之间和中外专利之间的分类相似性（公式 2），并分别构建论文的学科分类相似性网络和专利的 IPC 分类相似性网络。分类相似性值越小表示两个主题的学科 /IPC 分类差距越大，反之表示两个主题的学科 /IPC 分类越相近。

$$\rho_Y = \frac{\sum\left(R_Y - \overline{R_Y}\right)\left(R_{Y'} - \overline{R_{Y'}}\right)}{\sqrt{\sum\left(R_Y - \overline{R_Y}\right)^2 \sum\left(R_{Y'} - \overline{R_{Y'}}\right)^2}} \qquad (2)$$

Y 和 Y' 分别表示学科 /IPC 分类向量。然后，Y 和 Y'分别从小到大排序编秩。R_Y 和 $R_{Y'}$ 分别表示两个秩次。ρ_Y 表示两个聚类的分类相似性。

最后，从语义相似性网络中发掘相似性强的中外论文和专利聚类主题连接边，提取成熟方向；并进一步在学科分类网络和 IPC 分类网络中探查对技术空白发展甚至填补可能提供帮助的基础研究和技术应用方向。

4　实验结果及分析论证

本文以农业畜牧兽医基因组学的科技论文和专利为切入点，利用 Web of Science 平台收录的 SCI、SSCI 论文数据的摘要信息以及 Derwent Innovations Index 专利数据的摘要信息，对中外基因组学的基础研究和技术应用进行协同分析。其间，依据地址字段划分中国论文和外国论文，依据申请人国别代码字段划分中国专利和外国专利。检索时间范围为 2001 年至 2022 年。表 1 和表 2 给出了聚类结果。

4.1　技术空白发现

通过对比中外论文和专利的产出数量，可以发现，中国产出的畜牧兽医领域基因组学论文的数量约占全球的 1/3。专利的数量优势则更加明显。

通过分析中外论文的主题分布，可以发现中国基础研究基本完整覆盖本领域畜牧水产等相关物种和研究方向，并重点关注基因组测序和全基因组关联分析。具体来看：①在基因组测序方向，中外学者都产出了大量基础研究成果，内容涉及畜禽、水产和病毒测序等方向。②中国针对全基因组关联分析的基础研究主要关注畜禽、鱼类肠道、粪便微生物，以及与野生近缘物种的遗传多样性和重要性状基因挖掘。外国针对全基因组关联分析的基础研究主要围绕 3 个方向开展，其一是基因组学与遗传发育的关联研究；其二是通过基因组育种提高肉蛋奶质量和产量；其三是抗病分子育种。③外国对多维组学整合分析开展的研究比重较

表 1　畜牧兽医基因组学领域论文聚类主题列表

Table 1 List of clusters of paper topics in the field of animal husbandry and veterinary genomics

中国		外国	
类标签	主题名（类内节点数/个）	类标签	主题名（类内节点数/个）
4	使用 illumina 测序进行研究（452）	1	畜禽遗传标记研究（721）
6	水产物种基因序列研究（362）	8	鱼类基因组学与遗传发育研究（567）
1	畜禽全基因组关联研究（360）	7	多维组学整合研究（540）
5	禽类肠道微生物研究（255）	4	畜禽基因组测序（499）
2	畜禽和鱼类肠道微生物全基因组关联分析（240）	2	畜禽基因组育种提高肉质量和奶产量（448）
0	畜禽粪便微生物群落基因组研究（176）	9	畜禽肠道微生物研究（444）
9	禽类全基因组关联研究（160）	3	畜禽基因组育种用以抗病（439）
8	畜禽病毒基因组研究（156）	0	鱼类、昆虫全基因组测序研究（46）
7	畜禽 miRNA 序列研究（154）	6	畜禽病毒基因组测序（382）
3	家蚕基因组测序（144）	5	禽类基因组测序（246）

DOI：10.13998/j.cnki.issn1002-1248.23-0200

表 2 畜牧兽医基因组学领域专利聚类主题列表

Table 2 List of clusters of patent topics in the field of animal husbandry and veterinary genomics

中国		外国	
类标签	主题名（类内节点数/个）	类标签	主题名（类内节点数/个）
7	猪重要性状主效基因的基因型检测（234）	2	对畜禽基因进行整合、修饰等操作以预防疾病（33）
5	病毒基因检测（203）	0	重组病毒基因组的方法（30）
2	牛基因单核苷酸多态性检测方法（168）	8	核苷酸序列检测（25）
6	转基因和克隆胚胎方法（159）	9	畜禽与表型相关的基因型检测（17）
0	羊基因单核苷酸多态性检测方法（158）	5	标记、分离病毒基因组（16）
8	水产微卫星标记检测（154）	3	对细胞进行再生、分化和选择（12）
3	禽类基因多态性检测（142）	7	培育转基因禽类方法（8）
4	畜禽 PCR 扩增并检测扩增产物（140）	1	水产肌动蛋白基因和启动子的克隆（6）
1	分离基因组 DNA 技术（107）	6	家畜育种方法（3）
9	家蚕微孢子虫蛋白基因检测及利用（86）	4	其他（6）

大，虽然目前仅涉及基因组、转录组和蛋白质组的整合研究，但是已经比中国在这方面的基础研究先行一步。

中外专利技术布局主要围绕检测和转基因两个方向。其中，中国专利技术主要覆盖中国主要畜禽和水产品种，并重点集中在检测技术方向，在基因组相关的育种技术等方向上布局不足。虽然国外专利布局相对完整，但是申请数量较少，因此存在全而不满的局面。具体来看：①与检测相关的中国专利有 8 个主题，外国仅有 2 个主题；②与转基因相关的中国专利有 2 个主题，外国有 6 个主题，涉及基因整合、修饰、重组、分离、再生、选择等操作方法，并主要与疫病防控、病毒基因工程相关；③外国对家畜基因组选择的育种技术有相应的专利布局，但是中国面对该技术市场较为被动。

通过对比分析可知，中国在多个畜禽水产物种的转基因育种技术、多维组学整合分析技术等方向上还存在技术空白点。同时，在多维组学整合分析方向，虽然目前中国还没有形成有利的基础研究支撑，但是通过人才引进、前沿跟踪等其他方式，未来可能弥补这项研究空白。

4.2 技术空白预测

通过建立中外论文和专利聚类主题语义相似性网络，得到相似性最强的两条边，即中国论文主题 1- 外国论文主题 1- 中国专利主题 2- 外国专利主题 9 和中国论文主题 8- 外国论文主题 6- 中国专利主题 5- 外国专利主题 0，详见图 1。这两条边表征了当前该领域较为成熟的研究及技术方向。其中第一条边表征的主题内容是畜禽全基因组关联分析。从研究基础来看，该主题以基因和遗传学、农业、乳制品和动物科学等学科为支撑。同样由这些学科衍生出来的主题还包括外国论文主题 2（对畜禽基因进行整合、修饰等操作以预防疾病）、外国论文主题 5（标记、分离病毒基因组）、中国论文主题 9（家蚕微孢子虫蛋白基因检测及利用）、外国论文主题 7（培育转基因禽类方法）等，详见图 2。由于受到相同的基础研究支撑，全基因组关联分析与转基因、基因克隆等技术更容易在实验、方法等方面相互借鉴，甚至形成交叉融合。因此，未来转基因育种方向很可能与全基因组关联分析方向在基础研究领域形成融合。

从技术应用角度来看，畜禽全基因组关联分析这个主题涉及的技术主要应用在酶、核酸、微生物的测定、检验和制备方法等方面。在这些方面同样有应用的主题包括中国专利主题 8（水产微卫星标记检测）、中国专利主题 7（猪重要性状主效基因的基因型检测）、外国专利主题 6（家畜育种方法）、外国专利主题 1（水产肌动蛋白基因和启动子的克隆）等。由图 3 可知

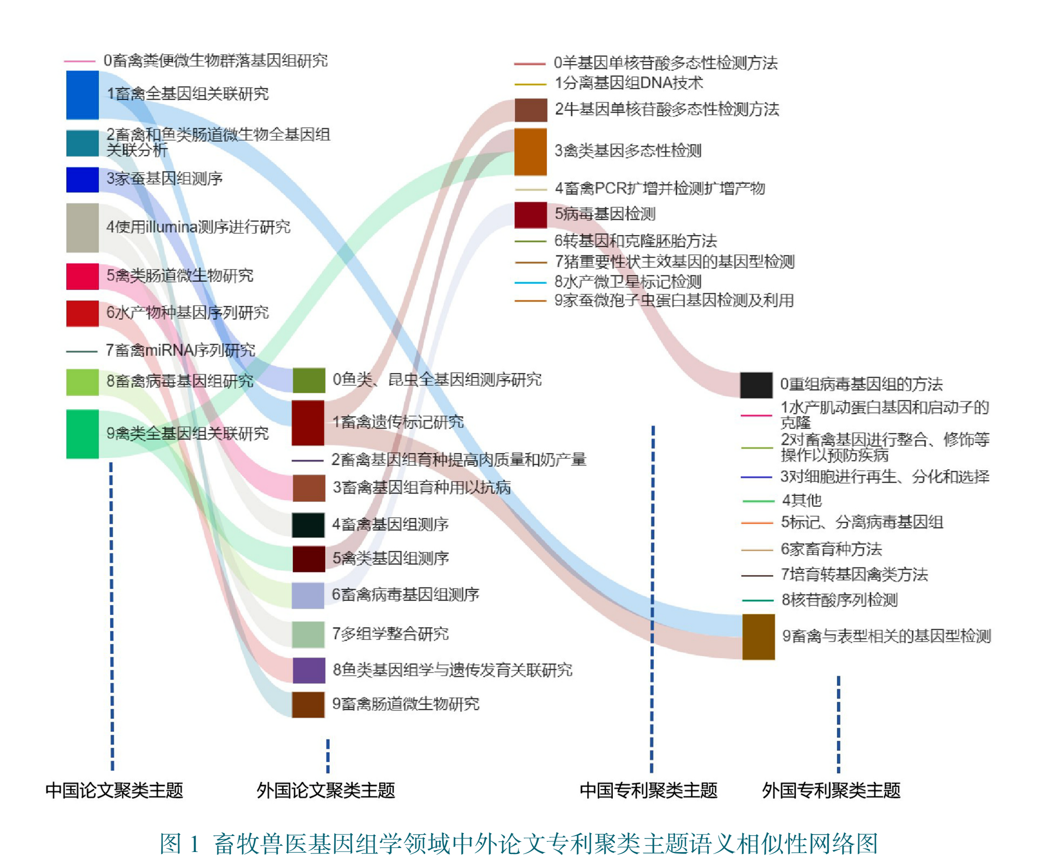


图 1　畜牧兽医基因组学领域中外论文专利聚类主题语义相似性网络图

Fig.1　Semantic similarity network of clusters between domestic and international papers and patents in the field of animal husbandry and veterinary genomics

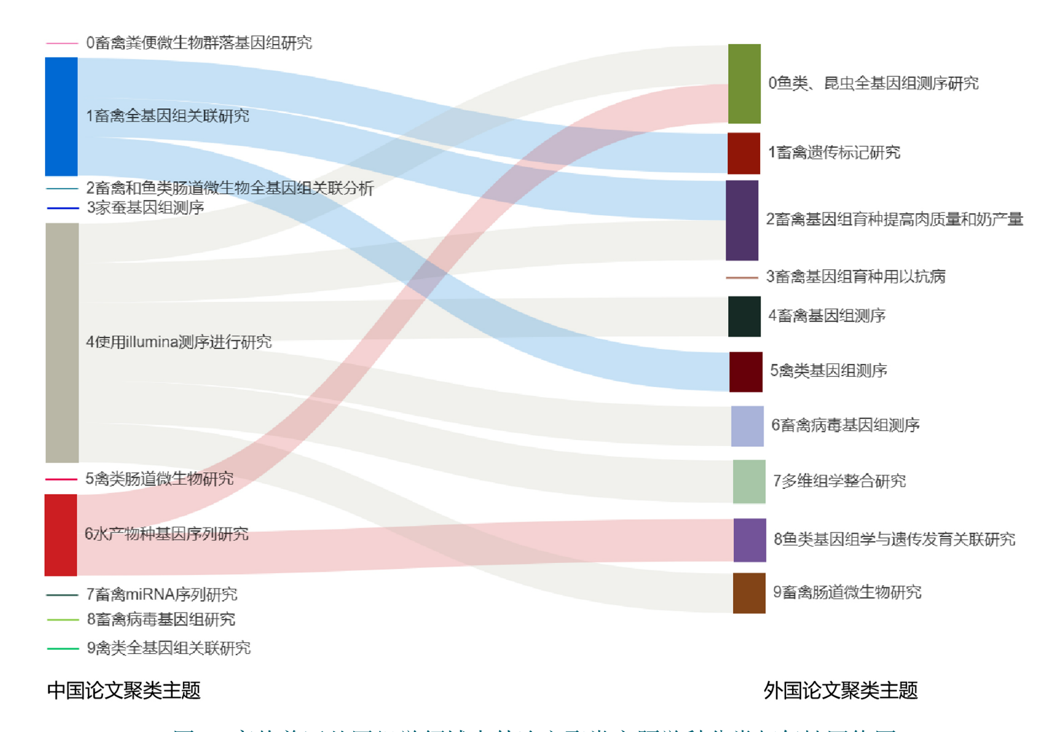


图 2　畜牧兽医基因组学领域中外论文聚类主题学科分类相似性网络图

Fig.2　Discipline classification similarity network of clusters between domestic and international papers in the field of animal husbandry and veterinary genomics

DOI：10.13998/j.cnki.issn1002-1248.23-0200

该技术应用方向还未形成较强的 IPC 分类相似性边。因此，虽然目前中国在转基因育种领域还没有形成独立的技术应用主题，但是通过人才引进等其他方式借鉴外国经验，并伴随转基因技术与全基因组关联分析在基础研究领域的进一步交叉融合，未来中国在转基因育种技术领域将大有发展，并有机会推动全基因组关联分析在育种领域获取更大应用空间。

第二条强相似性边表征的主题内容是畜禽病毒基因组分析。从研究基础来看，该主题由病毒学、兽医学衍生而来。同样由这两个学科衍生而来的主题包括外国论文主题 7（多组学整合研究）、中国论文主题 7（畜禽 miRNA 序列研究）、外国论文主题 9（畜禽肠道微生物研究）、中国论文主题 4（使用 illumina 测序进行研究）等，详见图 2。可见，当前病毒学和兽医学已经在多个研究方向上形成交叉融合。未来多组学整合分析很有可能首先与畜禽病毒基因组分析相融合，并相互推动继续深入发展。

从技术应用角度来看，畜禽病毒基因组分析主要应用在突变或遗传工程以及含有抗原或抗体的医药配制品方向。在这两个方向有相似技术应用的主题还包括中国专利主题 6（转基因克隆胚胎方法）、外国专利主题 5（标记、分离病毒基因组）等，详见图 3。可见，虽然目前多组学整合技术在全球范围内尚没有形成独立技术应用主题，但是通过与畜禽病毒基因组分析在基础研究方向的交叉融合，未来必将在畜禽病毒基因组技术领域大有可为，甚至可能促进转基因畜禽病毒多组学整合技术成为日后的技术研发热点。

4.3 结果验证

为了验证本文方法的有效性，对比了其他学者的相关研究成果。文章[25]通过科技论文分析识别出中外基因组测序、基因组拼接、数据库构建等热点。这些热点与本文聚类主题相呼应。同时，该文中对基因组学理论创新和整体性研究的预测与本文发现的多组学整合关联研究技术空白部分相互印证。文章[26]介绍了基因组学在发展中国家畜牧业的应用以及未来机遇。

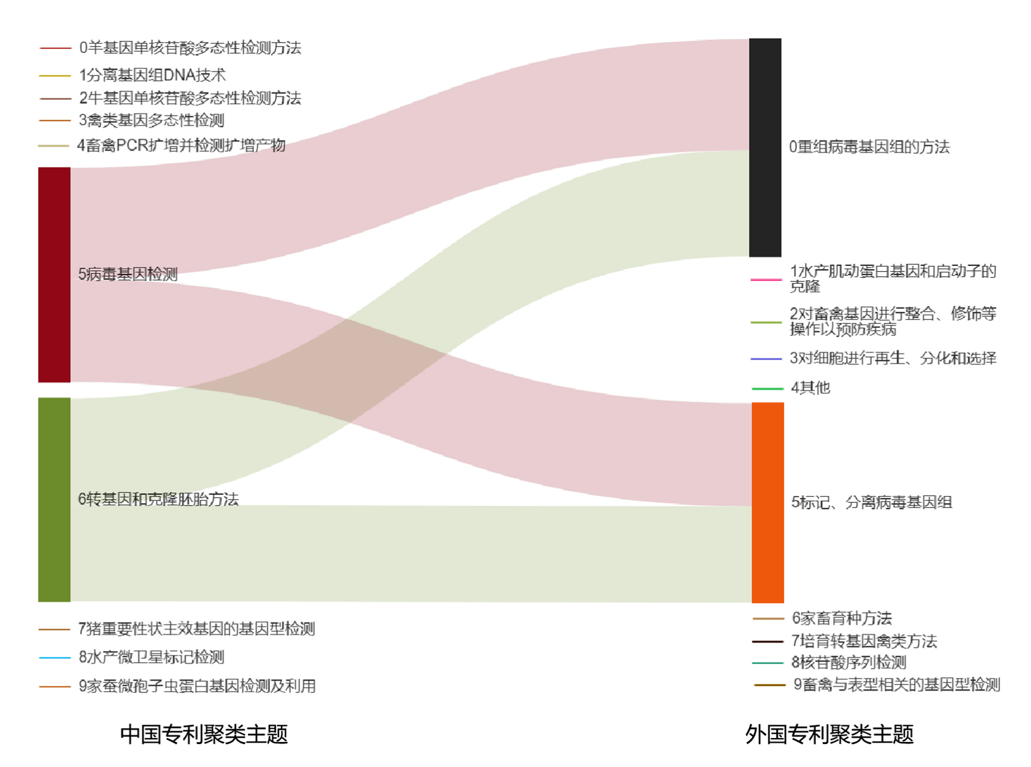


图 3 畜牧兽医基因组学领域中外专利聚类主题 IPC 分类相似性网络图

Fig.3 IPC classification similarity network of clusters between domestic and interna-tional patents in the field of animal husbandry and veterinary genomics

该文中讨论的全基因组关联研究、测序、病原体检测、疫苗开发和其他相关技术与本文方法发现的技术空白相吻合。前人的研究可以从侧面印证本文结论的可靠性和本文方法的有效性。另外本文识别的技术空白从中外论文和专利出发，数据源多样，数据基础更扎实，因此更具针对性和说服力。

5 结　语

本文针对畜牧兽医基因组学领域开展了中外技术空白对比研究，从中发现了中国技术空白，并对其未来发展给出了咨询建议。分析结果表明中国论文和专利产量大，但是技术架构布局没有外国完整，且中国论文的主题覆盖全于专利。具体来看，中国在多组学整合关联研究上缺少足够的基础研究支撑，技术条件也不完备；中国的转基因育种技术领域也属于技术空白。另外，转基因育种与全基因组关联分析、多组学整合与畜禽病毒基因组分析都存在进一步交叉融合的可能，未来将成为新的技术融合点。

本研究仍存在不足之处：①人为分析解读科技论文和技术专利的关联仍然耗费时间和人力。在未来研究中，会设计更自动的方法构建两种数据对象的关联对比方法。②专家解读聚类主题仍有可提升空间，未来可以考虑加入更多数据资料，增加标签信息，在减少人工标注工作的同时，为结果验证提供判断量化准确率的可能。

参考文献：

[1] 吴菲菲, 陈明, 黄鲁成. 基于 GTM 的 3D 生物打印专利技术空白点识别[J]. 情报杂志, 2015, 34(3): 58-64.

WU F F, CHEN M, HUANG L C. Identification of patent vacuums in 3D bioprinting based on GTM[J]. Journal of intelligence, 2015, 34(3): 58-64.

[2] 宫旭. 污水处理领域技术热点和技术空白点分析[J]. 中国化工贸易, 2017, 9(24): 91-93.

GONG X. Analysis of the technology hotspot and the technical gap in the technical field of wastewater treatment[J]. China chemical trade, 2017, 9(24): 91-93.

[3] 刘晓燕, 张淑伟, 单晓红. 技术融合的互补性与相似性研究[J]. 复杂系统与复杂性科学, 2023, 20(1): 81-87.

LIU X Y, ZHANG S W, SHAN X H. Research on complementarity and similarity of technology convergence[J]. Complex systems and complexity science, 2023, 20(1): 81-87.

[4] 葛小培. 专利地图的研究及其在生物医药领域中的应用[D]. 苏州: 苏州大学, 2010.

GE X P. Research of the patent map and its application in the biomedical field[D]. Suzhou: Soochow University, 2010.

[5] LEE S, YOON B, PARK Y. An approach to discovering new technology opportunities: Keyword-based patent map approach[J]. Technovation, 2009, 29(6/7): 481-497.

[6] YOON B U, YOON C B, PARK Y T. On the development and application of a self-organizing feature map-based patent map[J]. R&D management, 2002, 32(4): 291-300.

[7] CUI Y Y, ZHAO B B, XIA F. Technology opportunity recognition algorithm and decision assistance for non-drug antidepressant field in China[J]. Mathematical problems in engineering, 2022, 2022: 1-10.

[8] 王东兴, 王哲, 赵帆, 等. 氢燃料电池动力船舶技术标准现状分析与发展展望[J]. 交通信息与安全, 2023, 41(2): 157-167, 178.

WANG D X, WANG Z, ZHAO F, et al. State-of-the-art and prospect of technical standards for the ships powered by hydrogen fuel cells[J]. Journal of transport information and safety, 2023, 41(2): 157-167, 178.

[9] 赵磊, 李鹏, 李培根. 基于专利 IPC 共现分析的我国海水源热泵技术发展态势[J]. 制冷与空调(四川), 2023, 37(2): 320-325.

ZHAO L, LI P, LI P G. Development trend of seawater source heat pump technology in China based on patent IPC co-occurrence analysis[J]. Refrigeration & air conditioning, 2023, 37(2): 320-325.

[10] 罗恺, 袁晓东. 基于 LDA 主题模型与社会网络的专利技术融合趋势研究——以关节机器人为例[J]. 情报杂志, 2021, 40(3): 89-97.

LUO K, YUAN X D. A study on the technology convergence trend of patent based on LDA and social network - An example of joint robot[J]. Journal of intelligence, 2021, 40(3): 89-97.

DOI：10.13998/j.cnki.issn1002-1248.23-0200

[11] 吴晓燕, 胡雅敏, 陈方. 基于专利共类的技术融合分析框架研究——以合成生物学领域为例[J]. 情报理论与实践, 2021, 44(10): 179-184.
WU X Y, HU Y M, CHEN F. Research on technology convergence analysis framework based on patent co-classification: Taking synthetic biology as an example[J]. Information studies: Theory & application, 2021, 44(10): 179-184.
[12] 张紫芸, 王文发, 马乐荣, 等. 预训练文本摘要研究综述[J]. 延安大学学报(自然科学版), 2022, 41(1): 98-104.
ZHANG Z Y, WANG W F, MA L R, et al. A review of pre-training text summarization studies[J]. Journal of Yan'an university (natural science edition), 2022, 41(1): 98-104.
[13] MIHALCEA R,TARAU P. TextRank: Bringing Order into Texts[C]// Proceeding of the 2004 Conference on Empirical Method in Natural Language Processing, Association for Computational Linguistics. Special Interest Group on the Lexicon, 2004: 404-411.
[14] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality [C]// Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2. New York: ACM, 2013: 3111-3119.
[15] BOJANOWSKI P, GRAVE E, JOULIN A, et al. Enriching word vectors with subword information[J]. Transactions of the association for computational linguistics, 2017, 5: 135-146.
[16] REIMERS N, GUREVYCH I. Sentence-BERT: Sentence embeddings using Siamese BERT-networks [C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Stroudsburg, PA, USA: Association for Computational Linguistics, 2019: 3982-3992.
[17] KIROS R, ZHU Y K, SALAKHUTDINOV R, et al. Skip-thought vectors [J]. Advances in neural information processing systems, 2015, 28(1): 1-11.
[18] LOGESWARAN L, LEE H. An efficient framework for learning sentence representa-tions[J]. arXiv: 1803.02893, 2018.
[19] PAEA S, BAIRD R. Information architecture (IA): Using multidimensional scaling (MDS) and K-means clustering algorithm for analysis of card sorting data[J]. Journal of usability studies archive, 2018, 13: 138-157.
[20] ESTER M, KRIEGEL H P, SANDER J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise[C]// Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. New York: ACM, 1996: 226-231.
[21] 颜端武, 梅喜瑞, 杨雄飞, 等. 基于主题模型和词向量融合的微博文本主题聚类研究[J]. 现代情报, 2021, 41(10): 67-74.
YAN D W, MEI X R, YANG X F, et al. Research on microblog text topic clustering based on the fusion of topic model and word embedding[J]. Journal of modern information, 2021, 41(10): 67-74.
[22] BI DOKI M, MOOSAVI M R, FAKHRAHMAD M. A semantic approach to extractive multi-document summarization: Applying sentence expansion for tuning of conceptual densities[J]. Information processing & management, 2020, 57(6): 102341.
[23] DE MOURA VENTORIM I, LUCHI D, RODRIGUES A L, et al. BIRCHSCAN: A sampling method for applying DBSCAN to large datasets[J]. Expert systems with applica-tions, 2021, 184: 115518.
[24] ZHELEZNIAK V, SAVKOV A, SHEN A, et al. Correlation coefficients and semantic textual similarity[J]. arXiv: 1905.07790, 2019.
[25] 李旭彦, 朱正茂, 杨晓秋. 中外基因组学研究前沿的比较分析[J]. 中国基础科学, 2016, 18(3): 42-46, 50.
LI X Y, ZHU Z M, YANG X Q. Comparative analysis on the research fronts of genomics between China and foreign[J]. China basic science, 2016, 18(3): 42-46, 50.
[26] BOADU F, DU YF, XIE Y. Knowledge transfer received, entrepreneurial opportunity type, environmental dynamism, and innovative performance by overseas subsidiaries in China[J]. Technology analysis & strategic management, 2023, 35(3): 237-254.

Comparative Study on the Technology Gaps in the Field of Animal Husbandry and Veterinary Genomics between China and Foreign Countries

WU Lei[1,2], LI Xiaojie[3], DING Qian[1,2], SUN Wei[1,2], ZHOU Zhengkui[4]

(1. Agricultural Information Institute of CAAS, Beijing 100081; 2. Key Laboratory of Ag-ricultural Big Data Ministry of Agriculture and Rural Affairs.P.R.China, Beijing 100081; 3. Shenzhen Agricultural Genome Research Institute of CAAS, Shenzhen 518120; 4. Institute of Animal Science of CAAS, Beijing100193)

Abstract: [Purpose/Significance] In order to explore the technological gaps in Chinese im-portant agricultural fields and predict the future trends of these gaps, this study investigates technology opportunity discovery in the embryonic and developmental stages from the per-spectives of technology gap discovery and technology fusion opportunity discovery, provid-ing consultation and suggestions for decision-makers on the technology development op-portunities for technology innovation. [Method/Process] First, TextRank method was used to mine information in abstracts of papers and patents in this paper, which is a key sentence embedding method. The sentence vector clustering method was applied to extract topic sen-tences of papers and patents. Second, comparative analysis of topic clustering was used to detect technology gaps. Third, semantic similarity networks and classification similarity networks were used to discover the theme directions, which are likely to develop into cross-domain research areas with these technology gaps. [Results/Conclusions] The experi-mental results indicate that the proposed method can identify technological gaps. Combined with expert analysis, the experimental results can show the current development status and predict the trends of genomics technology in the field of animal husbandry and veterinary medicine. At the same time, this study can provide methodological and data support for genomics technology think tanks in the field of animal husbandry and veterinary medicine in China. Specifically, China has published a large number of papers and patents, but the tech-nical architecture layout is not as complete as foreign countries. The topics of Chinese papers are more complete than those of Chinese patents. In addition, China lacks sufficient basic research support in the integration and association of multi-omics, and the technical conditions are also incomplete. The field of genetically modified (GM) breeding technology is also recognized as a technological gap in China. In addition, it is possible that GM breeding and whole genome association analysis, multi-omics integration and viral genome analysis of livestock and poultry will become new technological fusion points in the future. There are still drawbacks in this study: It still takes time and manpower to manually analyze and interpret the relationship between scientific papers and technological patents. In the future research, more automated methods will be designed to construct correlation comparison methods between two data objects. Additionally, there is still room for improvement in expert interpretation of clustering themes. In the future, more data can be considered to add label information, reducing manual annotation work while providing the possibility of increasing quantitative accuracy in the result validation section.

Keywords: technology gaps discovery; key sentence extraction; sentence embedding clus-tering; genomics; intellectual property